DOCUMENT RESUME

ED 069 691 TM 002 142

AUTHOR	Passmore, David L.
TITLE	Objective Measurement in Occupational Education.
PUB DATE	May 72
NOTE	12p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Aptitude Tests; Individual Tests; *Mathematical Models; Measurement Techniques; Occupational Tests; *Predictive Ability (Testing); Statistical Analysis; Technical Reports; Test Construction; *Test Validity; *Vocational Education
IDENTIFIERS	Rasch Model

ABSTRACT

The model for educational measurement developed by George Rasch, a Danish psychometrician, is reviewed and its application to occupational educational testing discussed. The Rasch model is an adaptation from the theory of latent trait analysis. According to it, answering an item correctly is a function of the difficulty of the item and the ability of the person being tested. The raw scores serve as the basis for estimating the scale of ability. The author concludes that the properties of the Rasch analysis suggest solutions to a number of measurement problems in occupational engineering including developing and equating alternate forms of a test and estimating and interpreting changes in trainee performance. The item-free characteristics of this measurement model may allow the development of individually tailored tests. (DJ)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED 069691

OBJECTIVE MEASUREMENT IN OCCUPATIONAL EDUCATION

by

David L. Passmore
USOE Research Fellow
Minnesota Research Coordinating Unit for Vocational Education
University of Minnesota
Minneapolis, Minnesota

TM 002 142

May, 1972

OBJECTIVE MEASUREMENT IN OCCUPATIONAL EDUCATION

David L. Passmore[1]

There is an increasing tendency for employers and educators to make use of information accumulated by educational and psychological testing programs. In occupational education, there is a press for the development of means to measure and certify the competencies of teachers (Panitz & Olivo, 1970) as well as students (Baldwin, 1970; Ohio Trade and Industrial Education Services, 1970). Most test constructors in occupational education have chosen to be guided by classical measurement models even though these models may be inefficient in the light of recent advances in testing technology. The purpose of this paper is to review a promising model for educational measurement which was developed by George Rasch, a Danish psychometrician. This model has strong implications for future measurement practices in occupational education.

Nature of the Problem

Pervasive But Imprecise Nature of Measurement in Education

Educational and psychological tests are prevasive phenomena on the American educational scene. An estimated 250 million tests are administered yearly in the nation's schools (Brim, Glass, Neulinger, Firestone, & Lerner, 1969). Vocational counselors and educators have certainly accounted for a large portion of this test usage. Unfortunately, educational measurement is directed by relatively inexact guidelines when compared to measurement practices in Science as a whole. One reason given for this inexactness is that educational measurement practices lack the objective qualities often associated with the measurement of physical properties (Wright, 1968).

Degrees of Objective Measurement

Lacking objectivity. Suppose we are interested in estimating person A's tool recognition ability. It is often not feasible to directly observe this

ability. Instead, the magnitude of A's ability may be inferred from his responses to test items which compose Form A of a hypothetical XYZ Tool Recognition Test. An indicant of his ability is the sum of his item scores when each response is scored 0 (wrong) or 1 (right). It is useful to transform his total raw score into a value on a scale having a known distribution and uniform unit of measurement. However, commonly used transformations such as percentile ranks or standard scores may lack objectivity. Specifically, such scales tend to be sample-bound since an estimate of A's ability may change when he is compared to norms based on different occupational groups. Also, A's scale may be item-bound since his ability estimate may change if he is administered Form B instead of Form A of the test.

Possessing objectivity. Consider an example of the measurement of a physical property. The interpretation of the measurement of the length of an object such as a brick does not require limited reference to that specific set of objects used to calibrate our tape measure. Likewise, either a 12 inch or 36 inch tape measure may be chosen to estimate the length of the brick. The concern is that both tape measures have been calibrated in terms of some common scale to ensure comparable and generalizable results.

A Model For Objective Measurement

Criteria for Objectivity

The interpretation of a person's test score should be sample-free and item-free in order to achieve objective educational measurement. Rasch (1960) stated conditions for objectivity more formally:

(a) The scale used to convert raw test scores into useful estimates of ability must be generalizable to future test situations without requiring references to the characteristics of the sample used to calibrate this scale;

(b) Any number of different sets of test items selected from a pool of items which measure a unidimensional psychological attribute must yield similar estimates of a person's ability.

Rasch developed (1960) and explicated (1961, 1966a, 1966b) a mathematical model to meet these criteria. This model has been reviewed in depth by various measurement specialists (among others: Anderson, Kearney, & Everett, 1968; Birnbaum, 1968).

Rasch Model[2]

Theory.[3] The Rasch model is an adaptation from a probabalistic theory of test performance called latent trait analysis (cf. Lord, 1953). Individual differences in ability are assumed to exist on some unidimensional variable of interest. Test items, hypothesized to relate to this variable, are devised and then administered to a group of examinees. It is also assumed that these items are scored dichotomously (right or wrong) and that speed does not influence responses to the test items. Items are selected which, according to a statistical criterion, meet the assumptions of the model. Retained items are analyzed to develop a scale for transforming raw scores into ability estimates.

According to the model, answering an item correctly (O_{ni}) is a function of the difficulty (D_i) of the item and the ability (A_n) of the examinee or symbolically:

$$O_{ni} = f(D_i, A_n). \qquad (1)$$

When this relationship is summed over all items, examinees' raw scores are used to estimate the left side of equation (1). Rasch proposed that the raw score was a sufficient statistic for estimating the "latent" parameters on the right side of equation (1). Rasch's estimation procedure will not be elaborated here but it can be shown that the difficulty of each item and the ability estimate associated with each raw score do not change when they are derived using totally different groups of examinees.

This is not true in classical test theory. Essentially, the model provides a means of converting raw scores into estimates of ability. In addition, the scale of ability estimates has been shown to have the properties of a ratio scale. The importance of these features in a practical testing situation is shown in the example below.

Suppose that a hypothetical 60 item test of keypunching ability was administered to 100 keypunch trainees and that preliminary analyses revealed that 50 of these items might be scalable. Figure 1 is a raw score conversion table for this test. For example, if trainee X received a raw score of 25,

---

Insert Figure 1 Here

---

then his ability estimate would be 6.00 plus or minus, naturally, some amount of error. This is not markedly different from classical test score interpretation. However, the unique character of the Rasch technique has been theoretically and empirically supported. If responses from a new sample of trainees with somewhat different characteristics were analyzed, Rasch showed that the ability estimates derived for each raw score would be almost identical with those obtained in Figure 1. This means that the calibration of the ability scale would be freed from the quirks of the sample chosen for calibration purposes. Also, conversion tables similar to Figure 1 may be constructed for any number of subsets of the 50 items. Rasch showed that each examinee would receive a similar ability estimate on each of the tables even though these tables were based on different items. This means that ability estimation would be independent of the specific items chosen for measurement purposes. In addition, the ratio ability scale allows tremendous flexibility in the interpretation of results.

The mathematics encountered in using the Rasch model to infer ability from raw test scores are tedious. Fortunately, Wright and Panchapakesan (1969) have developed a computer program to determine the fit of the model to data and, also, to construct the raw score conversion table.

Empirical tests. The robustness of the Rasch model with respect to violations of kernel assumptions of the model has been supported by a number of computerized simulation studies (Brink, 1970; Noonan, 1969; Panchapakesan, 1969). Tests of the applicability of the model to demographic (Matthiessen, 1965), personality (Fowler & Bramble, 1972), and civil service (Durovic, 1970) data have produced encouraging results. The model has also been found adequate with achievement test data (Kearney, 1966; Tinsley, 1971). A study by Wright (1963) is outlined below to illustrate the type of research done on the Rasch model.

Wright administered the Law School Admissions Test (LSAT) to a sample of 1000 college students. The sample was divided at the median test score into two groups. Using a Rasch analysis on each group's item responses, two separate scales were independently calibrated for the LSAT. Although mean test performance of both groups differed greatly, the two resulting conversion scales were identical. Therefore, the sample-free claims for the Rasch technique were supported since such dissimilar groups yielded identical scales. Next, Wright separated the 48 item LSAT at the median item difficulty value into two sets of 24 items. The item-free character of the Rasch model was supported. Dissimilar sets of items produced similar ability estimates for all 1000 students.

Implications for Occupational Education

In the development of occupational competency exams on a national scale, it may be necessary to develop and equate alternate forms of a test. Angoff (1963) listed the multitude of technical problems associated with this task in traditional test analysis. But, using the Rasch technique, alternate forms

could easily be constructed since any set of a pool of calibrated items would produce identical information about examinees.

The ratio scale of measurement offered by the Rasch analysis may greatly facilitate the estimation and interpretation of change in trainee performance. Using ordinal or interval scales produced by classical test calibration, the amount of change exhibited by a trainee cannot be described simply as a ratio of his scale scores (Harris, 1963). However, an appropriate indicant of the magnitude of the difference between two ratio scaled values is merely the ratio of these two values (Stevens, 1951). Therefore, using a Rasch ability scale, a meaningful interpretation can be made of the change in a trainee's performance from point to point in time in a training program. Also, the magnitude of the difference between two trainees' ability may be easily estimated.

Baker, (1971, p.232) points out that prevalent test theory may be inadequate in light of recent and anticipated advances in educational technology. Individualized instruction in vocational education may also demand individualized attention in evaluation. Rigidly defined groups of examinees, times of testing and tests themselves may be inappropriate in the future. It may be necessary to tailor the test to the individual trainee who, for example, may have reading difficulties or other perceptual handicaps. The item-free character of the Rasch model would be very appealing in these circumstances. Using a pool of multimodal items calibrated by the Rasch technique in conjunction with tailored testing routines outlined by Lord (1971), the jump to a dynamic, individualized, and, possibly, computerized testing situation is not extreme.

Summary and Conclusions

The Rasch model represents a significant step in the refinement of objective educational measurement. The properties of the Rasch analysis suggest solutions to a number of measurement problems in occupational education including developing

and equating alternate forms of a test and, also, estimating and interpreting changes in trainee performance. The item-free characteristics of this measurement model may allow the development of individually tailored tests.

Enthusiasm for the advantages offered by the Rasch model must be tempered, however. The amount of research completed on this model is minute when compared to the wealth of theoretical and empirical work done on classical test theory. The applicability of the Rasch model must be more closely examined by applying the model in a wide range of practical situations.

References

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. The British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Angoff, W. H. Technical problems of obtaining equivalent scores on tests. Journal of Educational Measurement, 1964, 24, 11-13.

Baker, F. B. Automation of test scoring, reporting and analysis. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) Washington: American Council on Education, 1971.

Baldwin, T. S. The development of achievement measures for trade and technical education. Final Report, Washington, D.C.: Bureau of Research, U.S. Department of Health, Education, and Welfare, 1970.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. Part V of F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Brim, O. G., Glass, D. C., Neulinger, J., Firestone, I. J., & Lerner, S. C. American beliefs and attitudes about intelligence. New York: Russell Sage Foundation, 1969.

Brink, N. E. The effect of item discrimination and the range of item easiness estimates on the standard error of ability estimates using the Rasch model. Unpublished doctoral dissertation, University of California at Los Angeles, 1970.

Durovic, J. Application of the Rasch model to civil service testing. Albany, New York: New York State Department of Civil Service, 1970, ED 049 305.

Fowler, E. P. and Bramble, W. J. An analysis of personality data using a Rasch measurement model. A paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1972.

Harris, C. W. (Ed.) Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Kearney, G. E. The cognitive ability of Aboriginal Australians. Unpublished doctoral dissertation, University of Queensland, Australia, 1966.

Lord, F. M. The relation of test scores to the trait underlying the test. Educational and Psychological Measurement, 1953, 13, 517-548.

Lord, F. M. Robbins-Munro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31.

Matthiessen, P. C. Infant mortality in Denmark 1931-60. Copenhagen, Denmark: The Institute of Statistics, 1965.

Moonan, W. J. Evaluating trainee test performances by a G. Rasch measurement model: a dialogue. Paper presented at the 115h Annual Conference of the Militray Testing Association, New York, September 17, 1969.

Ohio Trade and Industrial Education Services. Trade and industrial education achievement test program. Columbus: The Ohio State University, 1970.

Panchapakesan, N. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Panitz, A. and Olivo, C. T. Report on the state of the art of occupational competency testing. New Brunswick, New Jersey: Department of Vocational-Technical Education, Rutgers University, 1970.

Rasch, G. Probabalistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Rasch, G. On general laws and the meaning of measurement in psychology. In Proceedings of the fourth Berkeley symposium on mathematical statistics. Berkeley: University of California Press, 1961, IV, 321-334.

Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), Readings in mathematical social science. Chicago: Science Research Associates, 1966a, pp. 89-108.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical Statistical Psychology, 1966b, 19 (1), 49-57.

Stevens, S. S. Mathematics, measurements, and psychophysics. In S. S. Stevens (Ed.), Handbook of experimental psychology. New York: Wiley, 1951, pp. 1-49.

Tinsley, H. E. A. An investigation of the Rasch simple logistic model for tests of intelligence or attainment. Unpublished doctoral dissertation, University of Minnesota, 1971.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 invitational conference on testing problems. Princeton, N. J.: Educational Testing Service, 1968, pp. 85-101.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Footnotes

[1]Mr. Passmore is a Research Fellow at the Minnesota Research Coordinating Unit for Vocational Education, University of Minnesota. Gratitude is expressed to the following University of Minnesota personnel for their critical review of this manuscript: F. Marion Asche, U. S.O. E. Research Fellow, Ronald McKeever, EPDA Fellow, and Dr. Cyril J. Hoyt, Professor of Educational Psychology.

[2]A bibliography of theoretical and empirical studies is available from the author upon request.

[3]This intuitive treatment may beg substantiation. The interested reader may find Rasch's publications less mathematically anemic.

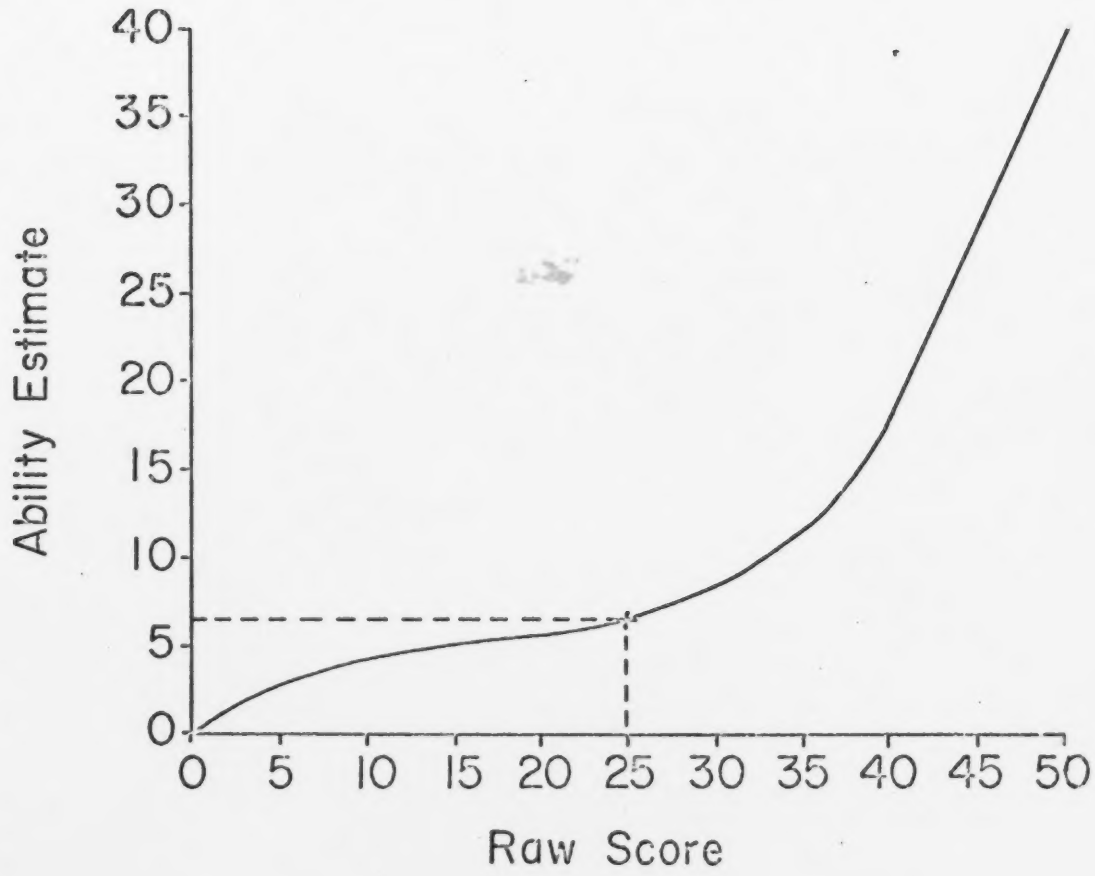

FIG. 1 A TYPICAL RASCH RAW SCORE CONVERSION TABLE.